A red and white logo

Description automatically generated

Course: ALY 6040(Data Mining Applications)

CRN:90331

Professor: Joseph Reilly

Boston payroll Dataset

Date: 15th-Aug-2023

Platform: Jupiter Notebook

**Why do I pursue this dataset?**

I selected this dataset not only out of curiosity but also as an opportunity to build a portfolio, showcasing my analytical skills and insights. This dataset presents a perfect chance for that.

Additionally, I recently watched a **Netflix series** called "**911**" and gained immense respect for the dedication of professionals like **police officers, firefighters, and doctors**. Observing the challenges, they face while working overtime inspired me to analyze how Boston's government plans workforce strategies for such crucial departments.

Another reason behind my choice is the **impact of job market changes in 2022**. I'm eager to uncover the specific effects of this shift on employment. Upon investigation, I learned that Boston's police, fire, and school departments play significant roles in the job landscape. Now, I'm intrigued by their earnings and whether job numbers have fluctuated over the years. This dataset is a treasure trove of information that can help me understand Boston's job dynamics and how these leading departments contribute.

Discovering that some individuals earn higher salaries was genuinely surprising. Perhaps my perspective has been influenced by common salary expectations in India.

**Why do these matters?**

Understanding how jobs change and how people work is important for knowing how a city's economy and society work. I'm focusing on Boston and using the payroll dataset to learn about the city's jobs. This dataset helps us figure out how employment in the city is changing and what things affect the kinds of jobs people have and how much they earn. From 2019 to 2022, there were big changes in the types of jobs people did, making this period interesting to study. As I was about to graduate and am searching for new jobs, so I think this dataset help me to understand and give me hope to not trust people's talks.

**Data Source**

We have a big collection of information about how much employees in different departments earn. There are 90926 rows and 15 columns in this dataset.

It tells us things like the names of employees, what departments they work in, their job titles, and how much they earn. This helps us understand how the city's workforce works.

I'm planning to focus on finding the top 3 departments and job titles from the dataset, and then separate the data based on that information. I'm not interested in using all the data because it's a lot to work with. Instead, I want to zoom in on the most important or popular departments and job titles.

I was using the last 4 years of data which are **2019,2020,2021,2022.**

**Variables**

We have the Following Variables:

* 'name': Employee's name.
* 'department': Employee's department.
* 'title': Employee's job title.
* 'Regular': Regular pay.
* 'retro': Retroactive pay.
* 'other': Miscellaneous/additional pay.
* 'overtime': Overtime pay.
* 'injured': Pay during injury/sick leave.
* 'detail': Detailed pay breakdown.
* 'Quinn': Quinn Bill-related pay.
* 'total': Total pay.
* 'zip': Employee's ZIP code

**What is Quinn Variable and Why do All Earners Have a master’s degree?**

The Quinn Variable is like a special thing we're looking at. It's connected to the Quinn Bill, **which is a program for police officers in Massachusetts**. This program gives officers extra money if they get higher education, like going to college.

**Now, why do all the people with the Quinn Variable have master's degrees?**

Well, the Quinn Bill is set up in a way that officers need to have master's degrees to get those extra bucks.

It's like a rule they made – if you want the bonus money, you need to go all the way to getting a master's degree.

Getting more money is nice, right? But not everyone needs a master's degree for it. Some police officers choose to get one, though. We're checking if many officers have masters and if it's changing over the years. Oh, and when prices go up a lot (inflation), people want to earn more money to keep up.

**Why am I only focusing on Top departments only?**

The reason I'm focusing on the top departments, jobs, or titles is to get a clearer picture of **what's happening in the city's workforce** without getting overwhelmed by too much information, and only top departments are contributing 70% of Boston's income. By identifying the most common or highest-impact departments and job titles, I can gain insights into the major trends and patterns in employment.

As the following fig. depicts the top departments with employee counts.

**A graph of a number of people

Description automatically generated with medium confidence**

In essence, by focusing on the top departments and job titles, I'm aiming to uncover the most crucial information that can give me a meaningful understanding of how employment is distributed and structured in the city.

**The top 3 Departments are as follows:**

* **Boston Public Schools (BPS):** This is the star department, but it comes with some variations. I'm aligning everything under one name – 'Boston Public School.'. Although I don’t have any plan to take a dig in this department ( As public facilities & services dept comes under the Boston school department)
* **Police Department:** I'm doing a detailed analysis here, tracking how different titles change over the years.
* **Fire Department**: Like the police, I'm examining changes in titles within the fire department.

**Were there any outliers or suspicious data?**

Based on the basic analysis I performed, I might want to investigate further exploratory data analysis to identify outliers or unusual patterns in the data. I can use box plots, histograms, or scatter plots to visualize the distribution of numerical variables and detect any potential outliers.

The boxplot shows the distribution of data for each variable. The box represents the interquartile range (IQR), and the whiskers extend to show the range of "normal" data points. Points beyond the whiskers are considered outliers.

* **Regular:** The data points are concentrated between 0.2 to 0.4, within the whiskers, indicating that most of the data falls within this range, with some variation.
* **Other:** There are outliers beyond the range of 1.0 to 1.2. These data points are unusual or different from the rest of the data in terms of their values.
* **Total:** The whisker extends up to the value of 1.0, also some have residual in 0.8 to 0. This suggests that the data is widely spread, and there are extreme outliers.
* **Injured:** The data show there are some outliers present in 0.6 & 0.8 but it’s not a lot.

**Business Questions**

Now, here's where the detective part comes in. I've got some cool questions that I want to find answers to:

* **Economic Journey:** How has the number of people living in Boston changed over the years? And what's the story behind how much money people make on average? This could tell us how the city is doing economically.
* **Neighborhood Secrets:** Can we figure out where rich and not-so-rich people live? It's like putting together a puzzle of where money hangs out.
* **Top Money Makers:** Which jobs make the most money? It's like finding out who the rockstars of the job world are.
* **Overtime Mysteries:** Do people, who make more money also work more hours? I'm curious if a bigger paycheck means more time on the job.
* **Paycheck Breakdown:** Let's look at the different parts of the paycheck—like extra pay, overtime, and when people are hurt at work. Maybe there's a pattern we didn't notice before.
* **The Quinn Bill Effect:** Lastly, I'm going to dig into this Quinn Bill thing. Does getting more education through this program help people's careers?

**Discovering the Gold:**

So, what's the big deal? By exploring all this data, I hope to uncover secrets about jobs in Boston. From understanding how the city's doing financially to finding out who's making the most money, this journey is like finding hidden treasure. And the best part? These treasures of information can help make Boston's future even better!

**Data cleanup**

1st, I import all the necessary libraries & create the File on my desktop & put all my 4 years' files. As the following code shows I put my path (the files I stored). The code “\*earnings, shows if u find any word which has earnings take that file and use it.

There are a lot of data-cleaning processes we need to do which I’ll do in the following steps.

1. Zip codes have very inconsistent values, so I transform them into 5 no. zip code, also there are a lot of NULL values in the zip code, so I filled in the NULL zip values using the most recent non-missing value encountered from the top.
2. We create a new variable “Year” from all files as each file has a year name, so we extract that and use it as a year.
3. **Co-relation analysis Findings:**

* The 'regular' and 'total' columns have the highest mean values, indicating they might be the major contributors to the overall earnings.
* The 'retro', 'other', 'overtime', 'injured', 'detail', 'Quinn', and 'zip' columns have lower mean values, indicating they may contribute less to the overall earnings.
* The 'regular' column has a considerable standard deviation, suggesting a wide variability in regular pay among employees.
* The 'retro', 'other', 'overtime', 'injured', 'detail', 'Quinn', 'total', and 'zip' columns also show notable standard deviations, indicating variability in these components as well.

**Cleaning the departments:**

In our dataset, we have a lot of titles and departments that fall under the same departments. We have many short forms that I want to clean. I am focusing on the Top 3 salary departments. I will map the specific titles to their respective departments. Additionally, I created a variable to save all the changes I was making.

The top 1st salaried department is 'Boston Public Schools', also known as 'BPS'. As we can see, many departments come under BPS, so we will rename them to 'Boston Public Schools' to maintain consistency.

**My primary objective** is to identify the unique departments for the years 2021 and 2022. As the dataset is extensive, I will focus on the data for the last two years. I want to compare the unique departments from the last two years and determine if any departments were present in both years or not. This analysis will help me understand the changes or consistency in department composition over time.

**Combining/consolidating the titles & departments:**

**Boston Police Department:**

The analysis of the 'Boston Police Department dataset reveals the most common police department titles before and after consolidation. Additionally, it provides a comparison of the unique title counts for the top 10 years in the dataset.

As the plots offer valuable insights into the impact of consolidation on title frequencies and the distribution of titles across different years. After data cleaning and consolidation, we observe a reduction in the counts of police officers, detectives, and police sergeants, indicating a streamlining of these positions.

On the other hand, titles such as call taker and clerk show minimal changes in the count, as they have relatively small numbers. However, when we focus on higher positions like police officers and detectives, we gain a clearer vision of the data after cleaning. Notably, in 2022, there is an increment in the number of unique titles, and the cleaned data provides a much clearer and more detailed view of the job landscape.

**Fire department:**

The fire department holds a crucial role in every city, and I deeply respect their contributions.

After performing data cleaning, we observe significant differences in the departments. The key titles within the fire department are firefighter & fire lieutenant, & their counts remain consistent across the years after cleaning. This indicates that the Boston Fire Department maintains a stable and steady structure in terms of titles, showcasing a sense of continuity and consistency in its workforce.

**Old vs Cleaned Titles & Departments visualizations**

After conducting thorough data cleaning, we can clearly observe its significance through the following visualizations. By comparing the original titles and departments with the newly cleaned and grouped ones, we can get to see new insights.

The graph illustrates that the number of job titles increased by approximately 1300 to 1500 in 2022. This suggests that even during times of recession, there are new job opportunities emerging. Additionally, the previous title counts show that we had over 1600 job titles in the market in 2022. If we had used this data without cleaning, we would have encountered more outliers, resulting in an inconsistent model.

Turning our attention to the departments, the visualization highlights the substantial impact of the cleaning process. Although we don't observe a significant increase in the number of departments, there is a notable difference when comparing 2022 to 2019, where we have 57 new unique departments. This indicates that some new departments were created or identified during the cleaning process, providing a more accurate and comprehensive representation of the data.

**Feature Engineering**

We created 2 new variables ‘average total earnings per capita’ & ‘average regular earnings per capita’.

* **Average total earnings per capita**: variable provides an average measure of total earnings per person, which can be useful for understanding the overall earnings distribution and trends within the dataset.
* **Average regular earnings per capita**: provides an average measure of regular earnings (excluding any additional earnings like overtime or bonuses) per person, which can offer insights into the compensation structure within the dataset.
* **Population Trend:** The population of Boston has been decreasing from around 700,000 in 2019 to approximately 660,000 in 2022. This downward trend indicates that the city's population is experiencing a decline over the years.
* **Average Wage Index (AWI) Trend:** The AWI in Boston has been increasing consistently over the years. In 2021, it was around 61,000, and in 2022, it rose to approximately 67,000. This upward trend suggests that the average wage level in the city has been growing.

The declining population might indicate various factors, such as migration patterns, changes in demographics, or economic conditions. It could mean that people are moving out of the city, The rising Average Wage Index signifies an improvement in wage levels for the employed population. This could be due to factors like economic growth, increased job opportunities, or wage growth in various sectors.

The total regular earnings for all individuals in the dataset in 2021 were around $1.55 billion, while the total earnings (including regular and other components) were approximately $1.87 billion. The data was collected for 22,546 individuals, and the average total earnings per capita were about $82,792.66. On average, everyone earned approximately $68,905.35 from regular earnings.

The total regular earnings for all individuals in the dataset in 2022 were around $1.60 billion, while the total earnings (including regular and other components) were approximately $1.93 billion. The data was collected for 23,204 individuals, and the average total earnings per capita were about $83,202.26. On average, everyone earned approximately $68,985.55 from regular earnings.

I want to find an average wedge index (awi)using the Boston population. I used awi and the population of the last 4 years from the available Boston data, which are 2019 to 2022.

Here's what each part of the output represents:

* **Bar Chart:** The blue bars represent the average earnings per employee for each year. This bar chart shows the trend in average earnings across the years 2017 to 2022.
* **Red Line:** The red line represents the trend of regular earnings over the years. It provides a visual representation of how regular earnings have changed from 2017 to 2022.
* **Yellow Line**: The yellow line represents the trend of the SSA average wage index over the years. The SSA average wage index is a measure of wage inflation, and this line shows how it has evolved from 2017 to 2022.

A graph with numbers and a red line

Description automatically generated

The visualization shows that regular earnings have experienced fluctuations over the years, with a notable increase in 2021 followed by a slight decline in 2022. In contrast, the SSA average wage index has remained relatively stable during the same period.

**Regression analysis**

We’ll create a regression model using overtime and regular variable.

The code filters the 'earnings' dataset to only include records from the year 2022 and where the 'overtime' earnings are greater than 1. Each data point in the plot represents an employee's earnings for the year 2022. The regression line on the plot helps visualize the correlation or relationship between 'regular' and 'overtime' earnings.

In summary, the plot allows us to examine the relation between regular and overtime earnings for employees in 2022, enabling us to identify any potential patterns or trends in the data that relate to these two variables. There are 6805 employees there.

**Why we’ll using regression analysis?**

A regression model was created to predict employees' likely to accept overtime based on regular and overtime earnings. The analysis found a positive correlation between regular and overtime earnings for employees with "normal" income levels.

We used Regression analysis because we wanted to understand how data changes in the independent variables are associated with changes in the dependent variable.

This helps us make data-driven decisions, understand trends, and predict outcomes in the context of the city's workforce and economic conditions. In summary, regression analysis allows us to model and quantify the relationships between different variables to gain valuable insights and support decision-making processes.

A graph showing a graph of injury

Description automatically generated

This visualization illustrates a trend among employees based on their earnings.

Individuals earning more than 150k tend to avoid working overtime.

On the other hand, those with earnings ranging from 50k to 150k are more inclined to take up overtime work.

So now we are creating a condition for people who like to do overtime. The following code shows the dataset select records where the sum of 'overtime' and 'regular' earnings is greater than 50,000, but less than 130,000. It also filters for records where 'overtime' earnings are greater than 6,000 and the ratio of 'overtime' to 'regular' earnings is between 0.25 and 1.5.

So, 25 % means, we’re going 2 compare regular income and overtime income if it’s lower than 25% it means they don’t like to do overtime or they don’t want to go for overtime & if it’s lower than 15 % means they like to do overtime (Human psychology often suggests that when individuals earn a sufficient income, they may be less inclined to work overtime, unless they hold higher positions)

**Understanding the Quinn Bill:**

Imagine a program that gives extra money to police officers who go to school. That's the Quinn Bill! It started in 1970 to encourage officers to learn more. By getting college degrees or higher, they get more money.

**Benefits of the Quinn Bill:**

**More Money:** When police officers get degrees, they get more money as a reward. This helps them earn a better living and stay updated with new skills.

**Level up a career**: The Quinn Bill can really help officers move up in their careers:

**Better Pay:** With more education, they can earn more because of the Quinn Bill. It also helps them learn about new technology.

**Skills Boost:** Learning more in school means officers know more about their job, like how to handle law matters. This makes them better at what they do.

**Getting Promoted**: In many police departments, having a college degree or higher can make officers eligible for promotions to higher positions. This means they can become leaders in the department.

**Better Chances:** When officers look for new jobs, having more education can help. It makes them stand out and look good to other police departments.

In short, the Quinn Bill is a way for police officers to get extra rewards for getting more education. It helps them earn more money, do better at their jobs, and move up in their careers.

**Impact on the Model**

The Quinn Bill's impact on police officers' career advancement is likely reflected in the income increase percentage after deducting Quinn's income. As we can see, employees are likely to pursue higher education. As I already attached a graph on the upper side, we can see that many people are not likely to pursue higher education. The percentage of employees not taking the initiative in the year 2019 was exceptionally high, but now it has decreased. Employees are more likely to take higher education. There are special departments that must be up to date, technology-wise, which are criminal departments.

Let's talk about data analysis. My father is a railway & criminal lawyer, and I noticed that the Governments often prioritize hiring individuals with higher degrees and advanced knowledge for various reasons:

Pursuing a higher degree in criminology plays a vital role in instances where some murder happens, and we have all the fingerprints and evidence. It's hard to crack the room condition most of the time; the police department forgets to collect important things or unintentionally neglects things that are important. This is where theoretical education plays an important role, whether we get a chance to learn new things and get a chance to work on case studies.

Overall, the analysis has provided valuable insights into various aspects of employee earnings, qualifications, and career advancement within the Boston Police Department.

Let’s see How many people pursue higher education.

**The salary distribution of top departments**

The graph illustrates the salary distribution among different departments. The Fire Department stands out as the highest-paid department, with salaries ranging between 70k to the highest at 130k. The Boston Police department maintains a consistent salary range between 90k and 100k. On the other hand, Boston Public Schools have the lowest salaries, starting at 40k. Notably, no department has salaries below 50k.

**Now we’ll want to see the employee who received Quinn's income and what kind of degree they pursued.**

In this code, I calculate the income increase percentage for each employee by comparing their regular and quinn incomes. We then categorize employees into 'Master,' 'Bachelor,' 'Associate,' or 'No Higher Degree' based on their income increase percentage (I got the information on the internet, If the income increase percentage is greater than or equal to 25%, the employee is classified as having a 'Master' degree. If it is between 20% and 25%, they are classified as having a 'Bachelor' degree. If it is between 10% and 20%, they are classified as having an 'Associate' degree. Any income increase percentage less than 10% results in the employee being classified as having 'No Higher Degree'.

The code shows the count of employees in each category for every year and plots a bar chart to visualize the number of employees with different degree qualifications over time. The Quinn bill might impact both quinn and regular incomes, influencing the degree qualifications of employees.

The stacked bar chart clearly illustrates the importance of pursuing a master’s degree. Employees with a master’s degree experience a noticeable increase in their pay salary compared to those without any higher degree. In 2022, around 10% of employees did not pursue any degree, while approximately 5% of employees in each year opted for a master’s degree. On the contrary, the number of employees without any higher degree has declined over the years. This trend indicates that having a master’s degree can significantly impact an employee's earning potential and may be a valuable investment in one's career growth.

In the next steps of my data analysis, I am focusing on zip codes to understand the distribution of high and low-income individuals. To accomplish this, I have installed the required libraries and downloaded Boston-specific zip code files and the national zip codes file, which contain latitude and longitude data.

In my code, I first read the zip code, latitude, and longitude data, as well as the earnings data for the year 2022. I then process this data to calculate the number of employees per zip code and merge it with Boston-specific zip code information. The 'Boston' and 'suburbs', contain relevant data that can be used to create various maps visualizing the distribution of employees in Boston and its suburbs based on zip codes.

Next, I will further segregate the employees based on their income levels into 'rich' and 'poor'. In the 'rich' category, employees have overall earnings of $150,000 or more, while in the 'poor' category, employees have 'total' earnings less than $60,000 and greater than or equal to $20,000.

The initial visualization of this data shows that a significant number of employees reside near the Dorchester area (zip code 1750), while a smaller number of people live near the Boston area.

There another analysis is the employer who’s earning good residing in the Dorchester area as Dorchester area are even don’t come under a safe area. It will give a good insight if we dig more into this co-relation.

A map of wealth and high elevation

Description automatically generated

**Interpretation of Results:**

1. **Correlation Analysis:**

The correlation analysis indicates that regular earnings have a strong positive correlation with total earnings, implying that regular pay is a significant driver of an employee's overall earnings. On the other hand, other earnings components, such as retro, overtime, injured, and detail, show weaker correlations, suggesting their relatively lesser impact on total earnings.

1. **Regression Analysis:**

The regression analysis demonstrates that the 'quinn' variable is a crucial predictor of 'regular' earnings, explaining more than half (54.6%) of the variability in regular earnings. This highlights the importance of considering the 'quinn' component in understanding and estimating an employee's regular pay.

1. **Degree Qualification Analysis:**

The analysis of income increase percentage and degree qualifications reveals that a substantial number of employees pursued master's degrees, followed by bachelor's and associate degrees. This suggests a possible correlation between education level and income increase, which could be an essential aspect to consider when assessing employees' career growth.

**Recommendations & Conclusions**

**Control Group Analysis:** Compare the regular earnings of employees who have benefited from the Quinn Bill with a control group of employees who have not pursued higher education degrees. This analysis can help quantify the specific impact of the Quinn Bill on regular earnings and provide insights into its effectiveness as an incentive program.

**Time-Series Analysis:** Conduct a time-series analysis to observe trends and changes in regular earnings before and after the implementation of the Quinn Bill. This can help evaluate the long-term impact of the program on employees' earnings and career growth.

The **Police Criminology department** presents an excellent opportunity for conducting a comprehensive analysis of the correlation between education pursued under the Quinn Bill and the frequency of overtime work among police officers.

Furthermore, the analysis can aid in identifying patterns or trends in overtime assignments based on the level of education attained by officers.

**Conclusion**

conducting an in-depth analysis of the relationship between education pursued under the Quinn Bill and overtime work can provide valuable insights that may lead to strategic improvements in the management of human resources and workforce planning within the Boston Police Department.

**Interpret the results of your analysis and explain what the results mean for the data owner.**

Let's say I am interpreting all these results to the Boston Mayor, as she's interested in getting to know all about Boston, like what's going on in Boston.

As the Boston Mayor, she’s eager to gain valuable insights into the city's workforce and employee earnings. She wants to understand the impact of the recession on the job market, identify the areas where job opportunities have decreased, and assess the economic position of the city. Additionally, she’s interested in understanding the qualifications of the city's employees and how we can encourage them to pursue higher education.

It is also crucial for me to determine the extent of overtime being worked by employees, especially in the police and fire departments, to ensure sufficient backup and resources in emergency situations.

My goal is to use data-driven analysis to make informed decisions that will promote the city's economic growth, enhance workforce engagement, and ensure the safety and well-being of our residents.

The data offers valuable insights for the Boston Mayor to make informed decisions. By focusing on competitive and fair regular pay, we can retain skilled employees. Supporting the Quinn Bill can maintain a well-educated police force. Investing in education leads to higher income for employees. Monitoring job trends allows strategic workforce planning. Tailored employee engagement can boost productivity. Population and economic data inform budgeting and resource allocation. Data-driven decision-making is the key to Boston's prosperity and safety.

The mayor can use these insights to lead ‘Boston’ toward a brighter future!

**References:**

* Turanga. (n.d.). Boston-Earnings-Analysis/0\_Boston\_Data\_Wrangling.ipynb at master · Turanga1/Boston-Earnings-Analysis. GitHub. <https://github.com/Turanga1/Boston-Earnings-Analysis/blob/master/0_Boston_Data_Wrangling.ipynb>
* Wheeler, A. (2021, March 15). Boston Department’s Average Earnings for Employees in Comparison to all Employee Earnings. Observable. <https://observablehq.com/@aliciawheeler19/boston-departments-average-earnings-for-employees-in-comp>
* Turanga. (n.d.). Boston-Earnings-Analysis/earnings.csv at master · Turanga1/Boston-Earnings-Analysis. GitHub. <https://github.com/Turanga1/Boston-Earnings-Analysis/blob/master/earnings.csv>
* Commonwealth of Massachusetts. (n.d.). Department of Developmental Services. Mass.gov. <https://www.mass.gov/orgs/department-of-developmental-services>
* Average Wage Index (AWI). (n.d.). <https://www.ssa.gov/oact/cola/awidevelop.html>
* Marytj. (2017). Chicago Payroll data visualization. Kaggle. <https://www.kaggle.com/code/marytj/chicago-payroll-data-visualization>